Bulletin of Toyohashi Sozo University
2011, No. 15, 37–46

運筆情報に基づいた古文書に含まれる仮名くずし文字の特徴分析

三　輪　多恵子
山　口　　満
島　田　大　助

デジタルアーカイブされた古文書の分析およびデータベース化を促進する目的で，計算機を用いた自動翻刻システムが求められている．本稿では，古文書に含まれる仮名つづき文字認識の基盤となる「仮名くずし文字の特徴分析」を行った結果について報告する．

キーワード：古文書，仮名くずし文字，運筆情報

I　はじめに

近年，様々な機関において，過去の映像や写真，文献等のデジタルアーカイブ化が進められている．アーカイブされた資料の効率的な利用方法のひとつとして，情報検索システムとの組み合わせが挙げられる．なお，一般的な情報検索システムでは“検索語”が使用されることが多く，文書（テキスト）以外の資料に対しては，検索タグ等の“メタデータ”の付与が必要とされている．

日本の古文書は，原本が世界各地に点在しており，デジタルアーカイブ化が進められている資料のひとつである．一方で，収集された古文書の多くがデジタル画像形式であり，収集と並行して“メタデータ”の作成という課題を抱えている．例えば，Web上で使用されるような全文検索の実現には，古文書本文のテキスト化が不可欠であり，現状では翻刻経験者の多大な労力が必要とされている．

この問題に対して，計算機を用いた自動翻刻システムの研究が行われている[1,2]．しかし，既存の研究はデータベース化された文字画像との比較によるものが中心であり，分析対象が漢字主体の定型文書に制限されているものが多い．古文書全文のテキスト化には仮名文字の個別認識が必要となるが，仮名特有の“くずし”や“つづき”により，画像比較による識別が難しく[3]，現時点で有効な解法は確立されていない．

筆者らは，仮名文字の筆順が字母（漢字）に依存する点に着目し，運筆座標情報（x, y座標値の変化）の直接比較による仮名つづき文字認識を試みた．その結果“つづき”の影響を軽減できることを確認し，一定の認識結果を得ている．その一方で，傾き，文字サイズ，くずしの程度，等の要因で認識率が大幅に低減する問題が報告されている[4,5]．

本稿では，上記の問題を回避するために，仮名文字の形状を特徴付ける成分の分析を行った結果について報告する．

Ⅱ　仮名くずし文字

仮名文字は，字母である漢字が簡略化される過程で定着した字体であり，“くずし”の程度により，その形状が多様に変化するという特徴がある．

図1に，くずし字用例辞典[6] に示された“仮名くずし文字”の例を示す．

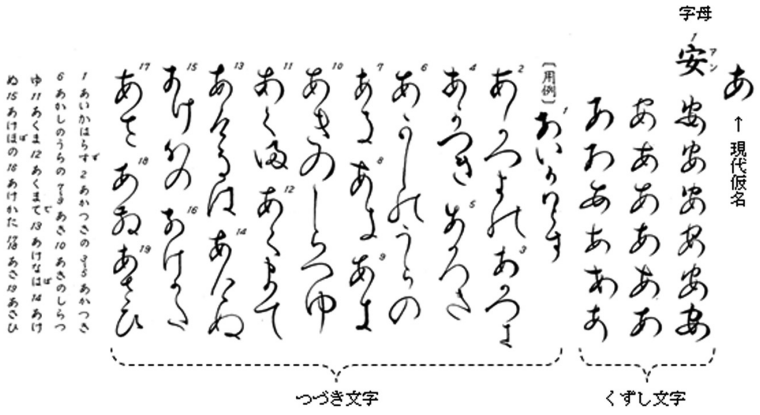


図1　“くずし文字”の例（字母：安）

用例辞典における“くずし”のパターンは，字母の形状（画数，等）によりその数に違いがあるものの，各仮名について約20種程度が挙げられている．今回の報告では，多数あるパターンから代表的な形状の2，3種類を抜粋し，特徴抽出に使用することとした．

本方式で使用した全ての“仮名くずし文字”画像を付録に示す．

1．運筆の座標系列

分析対象となる“仮名くずし文字”からの運筆情報の抽出について，以下に説明する．

今回の実験では，コンピュータディスプレイ上に表示した“仮名くずし文字”画像に対して，毛筆線のほぼ中央を通る運筆ラインを描画（ペン入力式のタブレット装置を用いて人手でトレース）することで，図2に示すように，時間に沿った運筆ラインのx, y座標値を取得する．なお，座標値は文字画像の左上を原点（0, 0），右下を最大値（256, 256）として取得しており，単位はピクセルである．

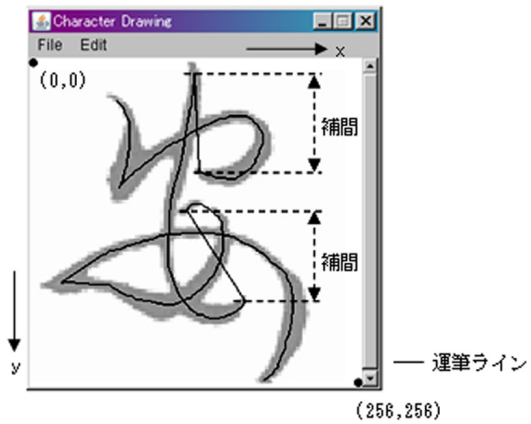


図2　運筆ラインの描画　—“あ”(字母：安) の例

また，1文字を1本の運筆ラインで表現するものとし，毛筆線が途切れている箇所については，以下の①，②のように処理を行うことで連続性を実現する．

① 墨のかすれ等を含む短い断線で，それ以前の筆跡方向の延長線上に次のラインの始端が存在しているような場合は，人為的な差異が生じにくいと考えられるため，手動で連続したラインをトレースする

② 断線箇所の前後で運筆ラインの属する字画が異なり，次のラインの始端が離れた箇所に存在するような場合は，手ぶれや揺らぎ等の人為的な差異が生じると考えられるため，断線箇所を線形補間する形でラインを自動生成する

例として，『の』から抽出したx, y座標系列を図3に示す．

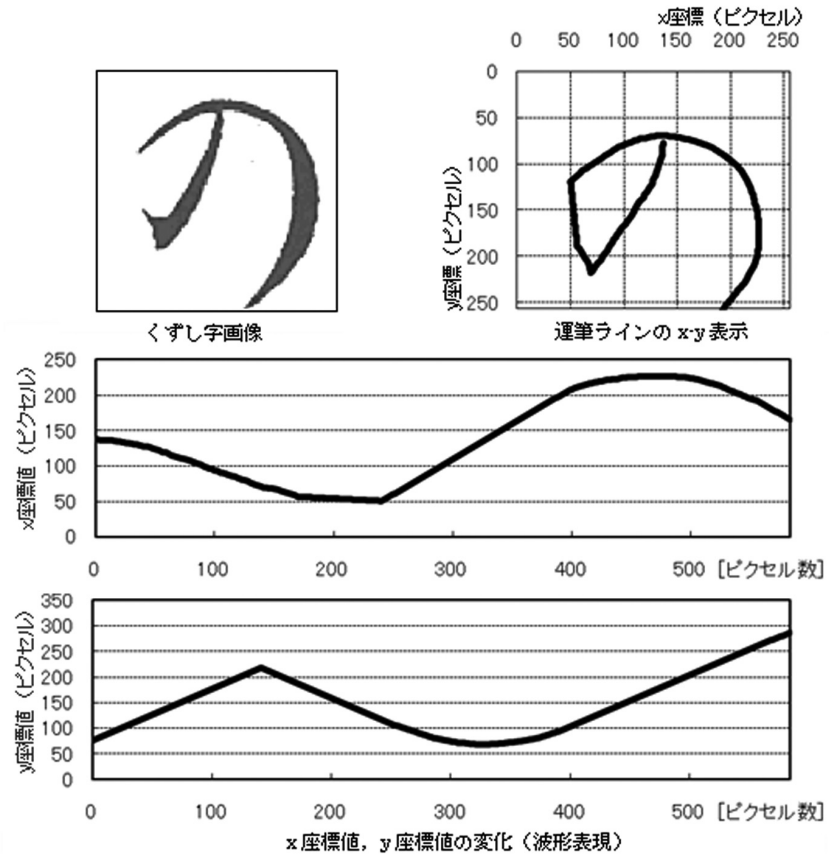


図3　くずし字からの運筆情報の抽出　―“の”(字母：乃) の例

2．運筆の方向（角度）系列

さらに，取得したx, y系列から運筆方向（角度）を算出する．

運筆系列上のiサンプル点における運筆方向θ_iは，垂直上方向を0［rad］として，時計周りに$0 \leq \theta_i < 2\pi$［rad］をとるものとし，図4に示すように，基準点（x_i, y_i）からTサンプル点離れた点（x_{i+T}, y_{i+T}）のなす角とした．なお，隣り合うサンプル点はx, y座標値がそれ

ぞれ$x, y-1 \leq x, y \leq x, y+1$の範囲に存在するため，2点のなす角が0，π/4，π/2，3π/4，π，5π/4，3π/2，7π/4［rad］の8方向に限定されてしまう．この問題を避ける目的でTサンプルの間隔を設けている．具体的なTの値については，次章3節において述べる．

例として，図3に示した『の』から算出した運筆角度系列 θ_iを図5に示す． $\theta_i \leq 2\pi$ となる箇所で不連続点が生じているものの，筆の進む角度が抽出できていることがわかる．

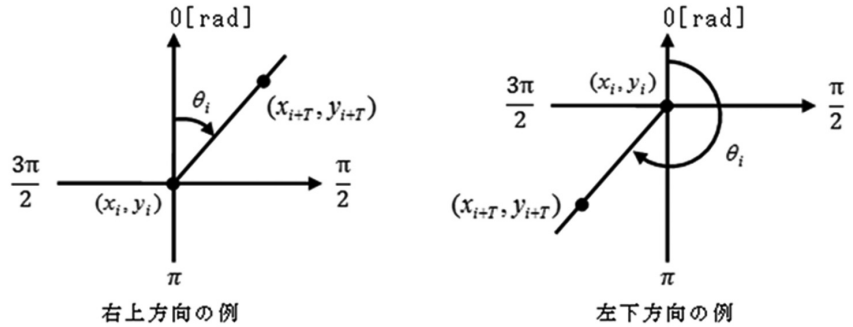


図4　運筆方向系列の θ_i

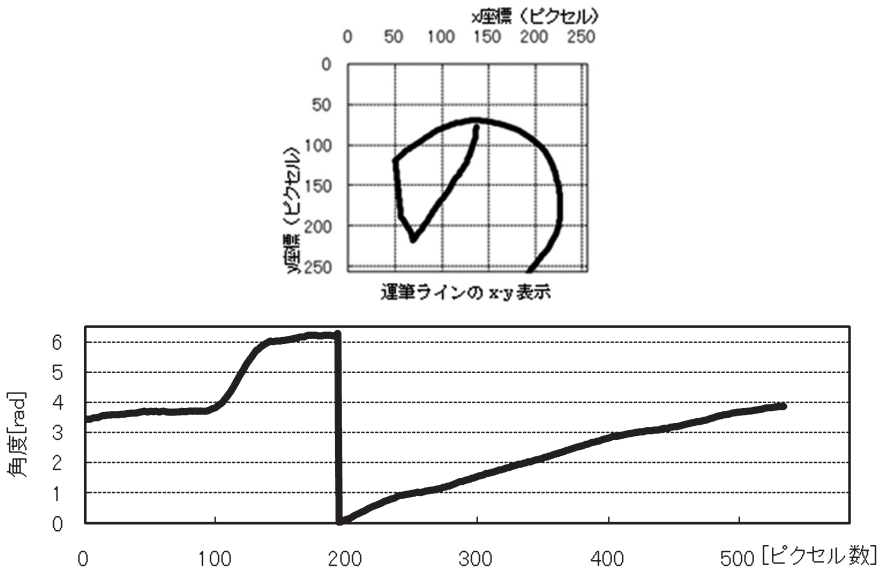


図5　くずし字からの運筆情報の抽出　―“の”(字母：乃) の例

Ⅲ　“仮名くずし文字”の特徴分析

Ⅱの1，2で取得した運筆座標系列x_i, y_i，運筆角度系列 θ_iの3系列を用いて“仮名くずし文字”の特徴分析を行う．

なお，今回着目した特徴は以下の3点である．上記3系列を用い，“仮名くずし文字”の分

類が可能か否かを調査することで，各特徴を用いた文字認識の可能性を確認する．

(1) 始点からの一定区間L_1における筆跡

文字の書き始めにあたり，現代仮名に置き換えた際の1画目の「入り」に相当する部分である．この部分は起筆と呼ばれており，翻刻経験者より，文字を判別する際に重要であるとの意見を得ているため，分析の対象として採用した．

(2) 終点からさかのぼった一定区間L_2における筆跡

文字の書き終わりにあたる部分であり，将来的に“仮名つづき文字”の認識を行う際に，運筆系列上で個々の文字を判別する（文字の終端を認識する）ために重要となると考えられるため，今回の分析の対象として採用した．

(3) 運筆ラインの角度変化

筆の進む方向（角度）を表しており，文字全体の形状を一定の基準で数値化したものとして捉えることができる．一方で，文字の形状によっては，運筆系列内において非常に複雑な変化を示すため，認識において有効な特徴成分の決定は非常に難しく，現時点で完全に整理できているとは言い難い．

今回の分析では，運筆ラインの角度変化から，筆跡の特徴が判別可能か否かに焦点を絞り，認識における角度変化の有効性について検証を行うことを主な目的とした．なお，着目した特徴は，手書き（右利き）における自然な動きに反する箇所（反時計回り）の筆跡である．

以下に，それぞれの特徴について“仮名くずし文字”を分類した結果について報告する．

1．始点からの一定区間L_1の筆跡

1）特徴判別

始点からの一定区間L_1について，表1に示す三つの性質に基づいた分類を検討する．

今回の分析では，区間L_1は文字画像の幅（256[px]）の約60%である150サンプルを用いた．また，“大きな増加”を識別するためのしきい値としてL_1の80%である120[px]を基準とした．なお，これらの値は，今回分析に使用した“仮名くずし文字”の運筆ラインから実験的に求めたものであり，最適な区間長さ，および，しきい値については今後さらなる検証が必要であると考えている．

表1　始点からの一定区間L_1における特徴

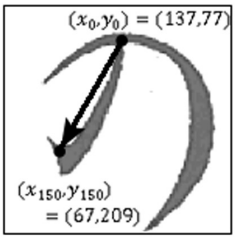
①	縦方向（下向き）	y座標値が大きく増加しており，かつ，L_1区間全体の運筆ラインのなす角度$\angle R_1$が垂直（下方向）に近い場合
②	横方向（左向き）	x座標値が大きく増加しており，かつ，L_1区間全体の運筆ラインのなす角度$\angle R_1$が水平（左から右方向）に近い場合
③	その他	上記2項目にあてはまらない場合

なお，運筆ラインのなす角度$\angle R_1$は，本来であれば最小二乗近似等を用いて傾きの傾向を求めることが望ましいが，計算量を削減するため，L_1区間内でy座標値（ないし，x座標値）が大きく増加しているという仮定に基づき，式（1）で近似している．

$$\angle R_1 = \tan^{-1}\left(\frac{y_{L_1} - y_0}{x_{L_1} - x_0}\right)$$

式（1）

例として，『の』からの角度$\angle R_1$の算出を図6に示す．なお，この『の』は，角度$\angle R_1$は垂直に近い値を示しているものの，$y_{L_1} - y_0 < 120$であるため③に分類されている．



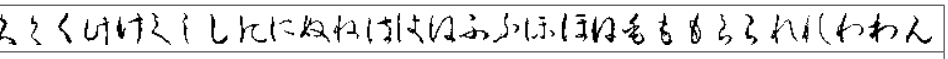
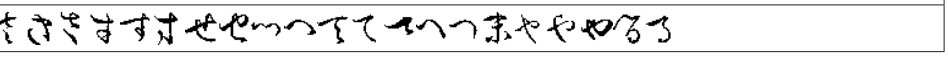
$$\angle R_1 = \tan^{-1}\left(\frac{y_{L_1} - y_0}{x_{L_1} - x_0}\right) = \tan^{-1}\left(\frac{209 - 77}{67 - 137}\right)$$
$$= 2.0583[\text{rad}]\ (117.93[\text{degree}])$$

図6　文字の書き始めの角度$\angle R_1$の算出　—“の”(字母：乃) の例

2）分類結果と考察

1）で述べた基準に従い，“仮名くずし文字”を3カテゴリに分類した結果を表2に示す．この結果から，今回用いた仮定により，人間が目視で確認できる特徴を正しく抽出していることが確認できた．

表2　始点からの一定区間L_1における特徴分類結果

①	
②	

なお，①下方向において完全に分類（検出）された仮名は『く』『し』『は』『ほ』『も』の5種類，②水平方向において完全に分類された仮名は『す』『て』『や』の3種類であり，それ以外の仮名については③その他，ないし，①と③，②と③に分かれて識別される結果となった．この原因としては，同一字母の“仮名くずし文字”であっても，“くずし”の度合いによって，①，②の特徴が表れにくいものがあることが考えられる．

一方で，今回分析を行った範囲では，同一字母において①と②に同時に分類される仮名パターンは出現せず，始点からの一定距離（区間L_1）における筆跡方向に基づいた分類①，②を用いて重複の無いカテゴリ分けの可能性が示せたと考えている．

2．終点からさかのぼった一定区間L_2における筆跡

1）特徴判別

終点からさかのぼった一定区間L_2について，表3に示す三つの性質に基づいた分類を行った．なお，区間L_2および“大きな増減”のしきい値については，実験的に求めた$L_2 = 120$，しきい値＝110をそれぞれ用いることとした．なお，前節と同様に最適な区間長，および，しきい値については今後さらなる検証が必要であると考えている．

表3．終点からさかのぼった一定区間L_2における特徴

①	縦方向（下向き）	y座標値が大きく増加しており，かつ，L_2区間全体の運筆ラインのなす角度$\angle R_2$が垂直（下方向）に近い場合
②	横方向（左向き）	x座標値が大きく増加しており，かつ，L_2区間全体の運筆ラインのなす角度$\angle R_2$が水平（左から右方向）に近い場合
③	その他	上記2項目にあてはまらない場合

なお，区間L_2における角度$\angle R_2$は，L_2区間内でy座標値（ないし，x座標値）が大きく増加しているという仮定に基づき，式（2）で近似している．ここで，Mは運筆系列長である．

$$\angle R_2 = \tan^{-1}\left(\frac{y_M - y_{M-L_2}}{x_M - x_{M-L_2}}\right) \qquad \text{式（2）}$$

例として，『の』からの角度$\angle R_2$の算出を図7に示す．

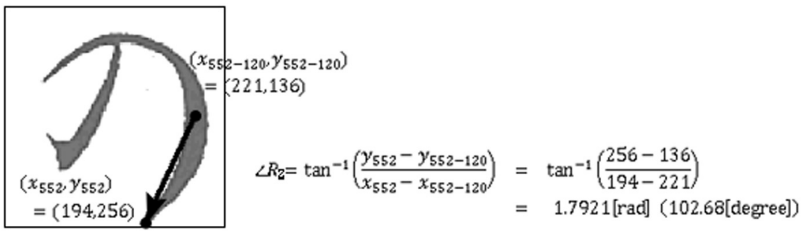
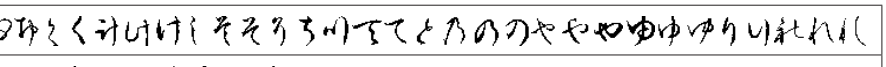


図7　文字の書き終わりの角度$\angle R_2$の算出　―“の”（字母：乃）の例

2）分類結果と考察

1）で述べた基準に従い，“仮名くずし文字”を3カテゴリに分類した結果を表4に示す．

表4．終点からさかのぼった一定区間L_2における特徴分類結果

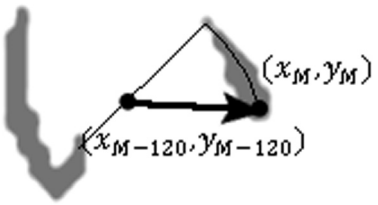
①	
②	

今回の分析では，①下方向において完全に分類（検出）された仮名は『け』『そ』『の』『ゆ』『り』『れ』の6種類であるのに対し，②水平方向において完全に分類された仮名は『ゑ』の1種類のみに留まった．これは，古文書が縦書きであることから，文字内における縦方向への変化が比較的安定しているためと推察される．

なお，今回の分析においては，複数の文字が目視による判別とは異なるカテゴリに分類された．この原因としては，

- 式（2）による近似的な角度算出において，最終画の長さを考慮していない
- “大きな増加”を判別するためのしきい値の設定が適切でない

の2点が考えられる．例として，視覚的には垂直方向変化に見える『い』が②に分類された理由について，図8で説明する．



$\angle R_2 \Rightarrow$ほぼ水平
$x_M - x_{M-120} > 110$

図８　終点からさかのぼった一定区間L_2における判別　―“い”(字母：以)の例

今回の分析においては，文字終端区間L_2における角度$\angle R_2$の算出には，終点(x_M, y_M)および終点からL_2点さかのぼった点（x_{M-120}, y_{M-120}）の２点を利用している．このため，図８に示すように，最終画が短い場合（ないし，短い区間で変化するような場合）は，その変化を無視する形になる．

このような場合，図８において矢印で示した終端部分の距離$|(x_M, y_M) - (x_{M-120}, y_{M-120})|$が比較的短くなることが予想されるため，$(x_M - x_{M-120})$ ないし $(y_M - y_{M-120})$ をしきい値と比較することで③に分類するように処理を行った．しかし，文字の形状によってはしきい値を超えてしまうことがあり，この『い』のように②として分類されてしまうことが判明し，しきい値による認識の難しさが明らかとなった．

なお，このような問題が明らかとなったものの，前述の７種類以外の仮名については，Ⅲの１で述べた変化と同様に，③その他，ないし，①と③，②と③に分かれて識別されているが，同一字母において①と②に同時に分類される仮名パターンは出現しなかった．このことから，終点からさかのぼった一定区間L_2における分類（表3①，②）を用いて重複の無いカテゴリ分けの可能性が示せたと考えている．

３．運筆ラインの角度変化

１）特徴判別

運筆ラインの始点から終点までの角度系列θ_iについて，以下の条件で分類を行った．

ここで，角度系列θ_iはⅡの２で述べた方法で算出しており，$i = 1$ ～系列長$- T$，$T = 30$として計算を行っている．なお，Tの値は実験的に求めており，最適な値については，今後さらなる検証が必要であると考えている．

表5①は，運筆ライン上のループ形状を反時計回りに辿る際に発生する状況であり，筆でくずし文字を描画する際には，右利きの人間にとっては自然ではない動きとなる．このため，文字仮名文字の形状を表す一つの特徴であると予想し，判別の対象とした．

なお，遷移発生の判別基準となるしきい値として$3\pi/4$［rad］を用いている．

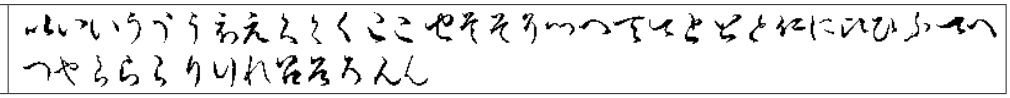
表5．角度系列θ_iに関する特徴

①	反時計周りのループ無し	角度系列中の２点θ_i，θ_{i+1}において，０から2πの遷移が発生しない場合

2）分類結果と考察

1）で述べた基準に従い，反時計周りのループが発生しない“仮名くずし文字”を検出した結果を表6に示す．

表6．角度系列 θ_i に関する特徴分類結果

①	

なお，表6①で完全に分類（検出）された仮名は『い』『う』『く』『こ』『し』『そ』『と』『へ』『ら』『ろ』の10種類である．なお，表6の結果において，一部，運筆の折り返し箇所に発生する極小ループによる識別率の低下が見られた．座標値系列と組み合わせることで，さらに分類の精度向上が見込めると考えている．

この結果から，仮名文字の形状を判別する際の角度系列の利用について，一定の有効性を示せたと判断できる．

4．考察

今回の特徴分析の結果から，運筆座標系列および角度系列に対するプログラム上の処理で，目視に近いカテゴリ分類が可能であることを確認した．対象となる仮名について，重複無くカテゴリ分類可能な特徴を発見できれば，段階的な判別を経て，文字認識が可能であると考えている．例えば，今回の検証においては，3①→1①と絞り込みを行うことで，『し』の可能性がある文字を漏れなく抽出することができる．一方で，今回検証を行った特徴（カテゴリ）だけでは，対象となる48種類の仮名を完全に分類するには不足しているため，さらに細分化するための特徴を見つけ出す必要があると考えている．

また，今回の検証においては，複数のカテゴリに分かれて分類されてしまう仮名や，目視とは異なる分類結果が得られてしまう仮名があり，分類のための条件（区間長，しきい値，等）については，さらなる検証が必要だと考えている．

Ⅳ　おわりに

今回の報告では，個別に切り出された“仮名くずし文字”の運筆座標系列および角度系列から，仮名文字の分類を試みた結果について述べた．前段研究における問題点である系列長の違いによる認識率の低下を解消する目的で，座標系列から角度系列を求めると共に，目視で判別できる特徴に対する自動判別の可能性について検証を行った．

実験結果より，始点および終点からの一定区間における運筆方向，運筆ライン上に発生するループ（反時計回り）の存在が仮名の分類に有効である事を示した．

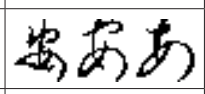
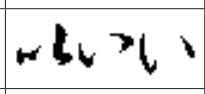
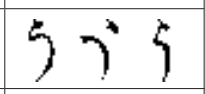
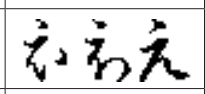
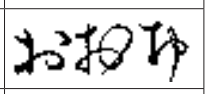
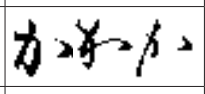
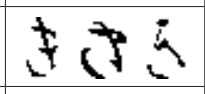
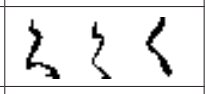
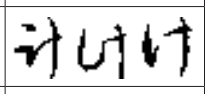
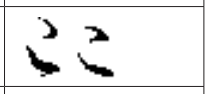
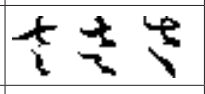
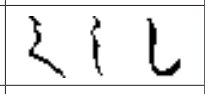
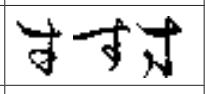
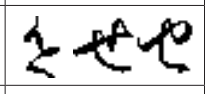
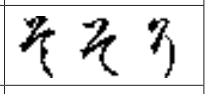
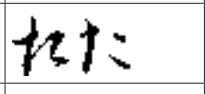
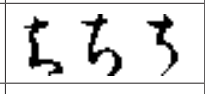
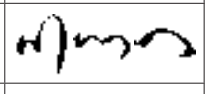
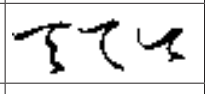
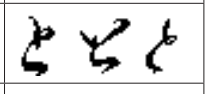
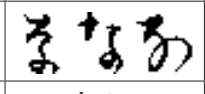
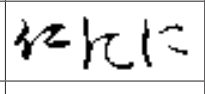
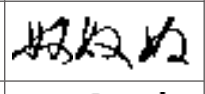
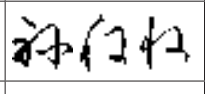
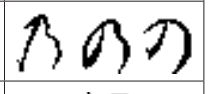
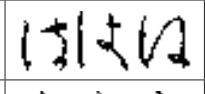
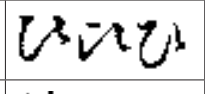
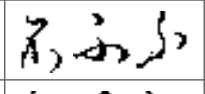
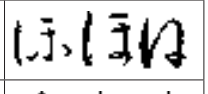
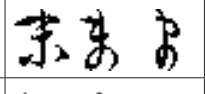
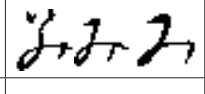
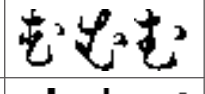
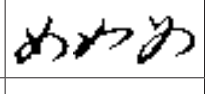
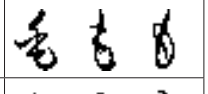
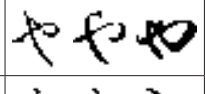
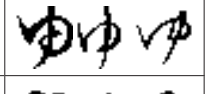
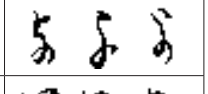
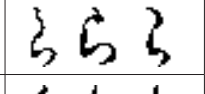
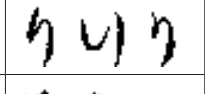
本研究の最終目標は，実際の古文書に含まれる仮名つづき文字の認識であり，今回の結果をさらに発展させ，各字母に対する様々な仮名形状を分類するための他の特徴について調査を進める予定である．

参考文献

［1］山田奨治，柴山守．「古文書を対象とした文字認識の研究」『情報処理』Vol.43，No.9（2002）pp. 950–955.
［2］山田奨治，柴山守．「n–gramによる古文書証文類翻刻支援の検討」『人文科学とコンピュータシンポジウム論文集』Vol.2000，No.17（2000）pp. 185–192.
［3］三好哲也，島田大助，三輪多恵子，舟久保登．「古文書における仮名文字認識に関する検討」『第19回ファジィシステムシンポジウム予稿集』（2003）pp. 269–270.
［4］三好哲也，島田大助，三輪多恵子，舟久保登．「運筆情報を利用した古文書におけるつづき文字認識」『電子情報通信学会ソサエティ大会講演論文集』（2006）A–4–7.
［5］舟久保登，三好哲也，島田大助，三輪多恵子．『江戸版本の読解を支援する運筆特徴を考慮したつづき文字の認識に関する研究』科学研究費（17500165）報告書（2008）.
［6］児玉幸多編．『くずし字用例辞典』東京堂出版（1993）.

付　録

表7．調査に使用した“仮名くずし文字”一覧

	a	i	u	e	o
A					
K					
S					
T					
N					
H					
M					
Y					
R					
W					
NN					